EPJ manuscript No.
(will be inserted by the editor)

Distance statistics in random media

High dimension and/or high neighborhood order cases

Cristiano Roberto Fabri Granzotti[1,a] and Alexandre Souto Martinez[1,2,b]

[1] Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP), Universidade de São Paulo (USP), Avenida Bandeirantes, 3900, 14040-901, Ribeirão Preto, São Paulo, Brazil
[2] National Institute of Science and Technology in Complex Systems (LNCT-SC), Universidade de São Paulo (USP), Avenida Bandeirantes, 3900, 14040-901, Ribeirão Preto, São Paulo, Brazil

Received: date / Revised version: date

Abstract. Consider an unlimited homogeneous medium disturbed by points generated via Poisson process. The neighborhood of a point plays an important role in spatial statistics problems. Here, we obtain analytically the distance statistics to kth nearest neighbor in a d-dimensional media. Next, we focus our attention in high dimensionality and high neighborhood order limits. High dimensionality makes distance distribution behavior as a delta sequence, with mean value equal to Cerf's conjecture. Distance statistics in high neighborhood order converges to a Gaussian distribution. The general distance statistics can be applied to detect departures from Poissonian point distribution hypotheses as proposed by Thompson and generalized here.

Key words. Poisson process – disordered media – random point problem

PACS. 02.50.-r Probability theory, stochastic processes, and statistics – 05.90.+m Other topics in statistical physics, thermodynamics, and nonlinear dynamical systems – 02.40.Dr Euclidean and projective geometries

1 Introduction

Consider a d-dimensional, unbounded, isotropic and homogeneous medium with disorder (points) generated by a Poisson process. The expected number of points in a volume V_d is $\lambda = \rho V_d$, where ρ is the point density. This disordered medium, although unlimited, can be represented computationally as a d-dimensional hypercube, containing N coordinates randomly distributed with uniform probability density function (pdf) along each edge (random point problem [1]). This is a possible way to construct a disordered medium, where, distances among the points are not fixed, but vary statistically. In this medium, it is possible to exploit the neighborhood and distance statistics.

Neighborhood statistics quantifies the probability of a point to be the mth nearest neighbor of its nth nearest neighbor. For $N \gg 1$, this probability was firstly calculated by Clark and Evans [2], for $m = n = 1$ and later generalized by Clark for mutual neighbors $m = n$ [3]. Dacey corrected expression obtained by Clark [4]. Neighborhood statistics was generalized by Cox, for $m \neq n$ [5] and interpreted in terms of multinomial distribution by Terçariol *et al.* [6].

Distance statistics quantifies the distance distribution of a given point to its kth nearest neighbor and can be applied in several disciplines. In Physics and Biology, this statistics can be used, for instance, in calculating the average separation between stars [7], aggregation in plant community [8,9], optimal tour for Euclidean salesman problem [10,11], Euclidean matching problem [12,13], partially self-avoiding walks [14,15], thin films [16], etc. In Computer Science, the nearest neighbor distance statistics can be employed as pattern classifier [17,18], other than being used to determine distance between network terminals [19], etc.

Up to now, only two aspects of this problem has been thoroughly addressed. The first one is the distance among points [20,21] and the moments of the highest order [22, 23] for different point distributions. The second one, is the distribution calculation for low-dimensional media, $d \leq 3$, for nearest neighbor [8,16] and arbitrary neighborhood [9]. The distance distribution to the kth nearest neighbor in an arbitrary dimension has been calculated by Martin [19], in the context of distance between internet access terminals. Despite the mathematical expression knowledge [19,24], the parameters influence on the distribution have not been

[a] e-mail: c.roberto.fg@usp.br
[b] e-mail: asmartinez@ffclrp.usp.br

fully addressed, mainly in cases of high dimension and neighborhood order.

These limiting cases are non-trivial, because of the function ratio $\Gamma(z+x)/\Gamma(z)$, for $z \gg x$, where $\Gamma(z)$ is the gamma function. If one considers simpler expansion, inconsistencies like undefined central moments, such variance and skewness, occur. Our main contribution is to correct these inconsistencies using higher order terms in this ratio expansion. We calculate the distribution in these limiting cases and proof important results as the ones conjectured by Cerf *et al.* [10].

In this paper, we obtain the analytical expressions for high dimensionality distance statistics that leads to an equivalence of the random link model. Also, the high neighborhood order is addressed, the resultant distribution converges to a Gaussian due to central limit theorem. The distance statistics can be used not only for predicting separation between neighbors, but also to detect departures from Poissonian hypothesis, as proposed by Thompson [9] and generalized here. Furthermore, we expand special cases of distance statistics varying dimensionality and neighborhood order.

Our paper is organized as follows. In Sec. 2, we obtain the pdf of distance statistics in two distinct ways: throughout geometric interpretations and cumulative functions. This pdf is described by the generalized gamma distribution [25,26]. In Sec. 3, we calculate the high neighborhood order $k \gg 1$, and high dimension $d \gg 1$ limiting cases. In this way, we demonstrated mathematically the Cerf's *et al.* conjecture [10], and consider the combination of the limiting cases. In Sec. 4, we explore special cases of distance statistics by varying dimension and neighborhood order and propose a generalized hypothesis test to quantify deviations from Poissonian spatial process. Finally, on Sec. 5, we present the conclusions.

2 Statistics of the random point problem

In this section, we obtain the analytical distance distribution expression and validate it by Monte Carlo simulations. In addition, we collapse the data with nearest neighbor distance distribution. Consider a d-dimensional medium, with density ρ, where $\rho = \rho_1^d$ and ρ_1 is the one-dimensional medium density. The previous argument keeps the mean distance among points constant, which allow us to compare different system dimensionalities. The expected number of points in a hypersphere of radius l is $\lambda = N_u l^d$, where $N_u = \rho\pi^{d/2}/\Gamma(1+d/2)$ is the number of points in a d-dimensional sphere with unitary radius. The probability of k points to fall into a sphere of radius l is given by the Poisson formula, $P(k) = e^{-\lambda}\lambda^k/k!$.

The first method to derive the distance statistics is based on geometric arguments. The probability of k points to fall inside a sphere of radius $l+dl$ is written as product of the probability of a sphere of radius l to contain $k-1$ points and the probability of the spherical shell thickness dl to contain only one point: $f^{(k)}_{\rho,d}(l)dl = P(k-1)P(1)$. As $dl \ll l$, the probability density function becomes:

$$f^{(k)}_{\rho,d}(l) = \frac{dN_u^k l^{dk-1}}{\Gamma(k)} \exp(-N_u l^d), \tag{1}$$

where k is neighborhood order and can be mapped on the generalized gamma distribution:

$$\mathrm{GG}(x|\theta,k,\beta) = \frac{1}{\beta\theta\Gamma(k)}\left(\frac{x}{\theta}\right)^{k/\beta-1} \exp\left[-\left(\frac{x}{\theta}\right)^{1/\beta}\right], \tag{2}$$

with: $\beta = 1/d$ and $\theta = N_u^{-\beta} = [\rho\pi^{d/2}/\Gamma(1+d/2)]^{-\beta}$, which depends on the point density and medium dimension. It is non-trivially affected by medium symmetry. If one considers a computer simulations, θ is only affected by media boundaries through out ρ. If, in one hand, one considers a d-dimensional hypercube with edge length $\mathcal{L}$ and N points, then $\rho = N/\mathcal{L}^d$. In the other hand, if one considers a sphere: $\rho = N\Gamma(1+d/2)/\pi^{d/2}\mathcal{L}^d$.

Monte Carlo simulations validated Eq. 1. The medium consisted of a cube with N points and density $\rho = N/\mathcal{L}^d$. The results of Eq. 1 applied to this limited medium is an approximation due to boundary effect, since points near the boundaries have fewer neighbors. Periodic boundary conditions minimize this effect. This validation is depicted in Fig. 1, where one sees that increasing the neighborhood order, statistics distribution become more symmetric around their mean value. Moreover, the numerical experiments consider a finite number of points. The correction for finite size system is of order $1/N$ for the mean distance [20]. Further, Eq. 1 in terms of $\lambda = N_u l^d$, number of points in d-dimensional sphere of radius l, is collapsed into $f^{(k)}(\lambda) = \lambda^{k-1}e^{-\lambda}/\Gamma(k)$.

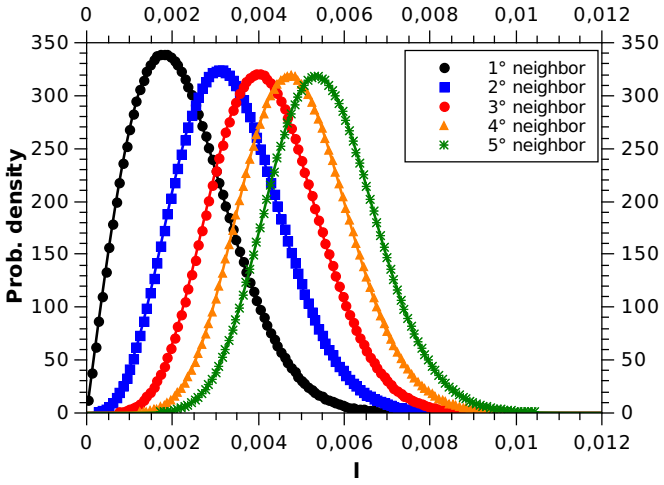

Fig. 1. Comparison of the analytical results generated by Eq. 1 (full lines) up to the fifth neighbor in a two dimension medium with Monte Carlo simulation. Simulations parameters are $\rho = 50000$, with periodic boundary conditions. The increase of neighbor order makes distance statistics more symmetric with respect to the mean value.

The second method is based on the cumulative distribution function. Consider firstly a point i and its nearest neighbor, in two-dimensional medium. The probability of not finding any other point closer than l is $P(k=0) = e^{-\rho\pi l^2}$. The random variable L, that describes the distance up to point i nearest neighbor, $L > l$ if there is no points in area πl^2, consequently $P(L > l) = e^{-\rho\pi l^2}$. Thus, L cumulative distribution function is: $P(L \leq l) = 1 - P(L > l) = F^{(1)}_{\rho,d}(l)$. The pdf that describes the distance to first neighbor is the $dF^{(1)}_{\rho,d}(l)/dl$: $f^{(1)}_{\rho,d}(l) = 2\rho\pi l e^{-\rho\pi l^2}$. This reasoning can be extended to arbitrary neighborhood and dimensionality, and leads to Eq. 1.

The mean distance of a point i to its kth nearest neighbor is:

$$\langle l^{(k)}_{\rho,d}\rangle = N_u^{-\beta}\frac{\Gamma(k+\beta)}{\Gamma(k)} \tag{3}$$

which has been firstly obtained by Percus e Martin [20], and factorizes in density and neighborhood order. The $l^{(k)}_{\rho,d}$ variance is:

$$\sigma^2(l)_{\rho,d,k} = N_u^{-2\beta}\left[\frac{\Gamma(k+2\beta)}{\Gamma(k)} - \left(\frac{\Gamma(k+\beta)}{\Gamma(k)}\right)^2\right], \tag{4}$$

which is difficult to analyze, because of the $\beta = 1/d$ ratio in the gamma function argument. When $k \gg \beta$, one can consider a simple expansion of ratio $\Gamma(k+\beta)/\Gamma(k)$ as k^β or $k^\beta e^{-\beta/k}$, however the calculation of central moments as variance and skewness are inconsistent. Expanding to higher orders, for $z \gg x$, one has:

$$\frac{\Gamma(z+x)}{\Gamma(z)} \approx z^x \exp\left(\frac{-x}{2z} + \frac{3x^2}{4z}\right). \tag{5}$$

According to Eq. 5, the mean and standart deviation of $l^{(k)}_{\rho,d}$ can be approximated to: $\langle l^{(k)}_{\rho,d}\rangle \approx N_u^{-\beta}k^\beta$ and $\sigma(l)_{\rho,d,k} \approx c\beta N_u^{-\beta}k^{\beta-1/2}$, with $c = 3/2$, indicating that the mean distance, in high dimensionality, is weakly affected by the neighborhood order, while the variance decays very rapidly. This occurs because the volume of a sphere is almost concentrated in a very thin spherical shell, when $d \gg 1$. The skewness of Eq. 1 depends non-trivially on k and β

$$\gamma_1 = \frac{2 - \Omega_1^2(k,\beta)/\Omega_2(k,\beta) + \Omega_1^3(k,\beta)/\Omega_3(k,\beta)}{(1 - \Omega_1^2(k,\beta)/\Omega_2(k,\beta))^{3/2}}, \tag{6}$$

where $\Omega_n(k,\beta) = B(k,n\beta)/\Gamma(n\beta)$ and $B(a,b) = \Gamma(a)\,\Gamma(b)/\Gamma(a+b)$ is beta function. The skewness is modified only by neighborhood order and dimension, being independent of medium boundaries and density, using Eq. 5 it can be approximated to $\gamma_1 \approx 6\beta k^{-1/2}$. This simplification accurately describes the skewness behavior around the mean value due to the neighborhood order, see Fig. 1. Also, the skewness factorizes in neighborhood order and dimensionality.

3 Limiting cases

In this section, we analyze the behavior of Eq. 1, firstly in the limit $d \gg 1$, next for $k \gg 1$ and finally both limits simultaneously. Although straightforward, these calculations present some pitfalls which are properly stressed.

3.1 High dimensionality

Let us introduce a new variable $y = (l - \langle l^{(1)}_{\rho,d}\rangle)/\sigma_{\rho,d,1}$, which standardizes the distance by the mean separation between the points. As $d \gg 1$, one has $\langle l^{(1)}_{\rho,d}\rangle \approx N_u^{-\beta}$ and $\sigma_{\rho,d,1}(l) \approx c\beta N_u^{-\beta}$. Distance can be rewritten as follows

$$l = N_u^{-\beta}(1 + \beta c y). \tag{7}$$

with $c = 3/2$ and $\beta = 1/d$. In the y variable, the pdf is obtained from the application of probability transformation law to Eq. 1, using l from Eq. 7. Starting with $k = 1$, one finds the Gumbel distribution $(1/|\bar\lambda|)\exp[-(x/\bar\lambda) - \exp(-x/\bar\lambda)])$, with $\bar\lambda = -1/c$:

$$g(y) = c\exp[cy - \exp(cy)], \tag{8}$$

which describes the minimum deviation from the expected separation: $\langle l^{(1)}_{\rho,d}\rangle = N_u^{-\beta}$. For higher neighborhood orders, one has:

$$g^{(k)}(y) = \frac{c}{\Gamma(k)}\exp[cky - \exp(cy)], \tag{9}$$

which is the log-gamma distribution $([1/|\bar\lambda|\Gamma(k)]\exp[kx/\bar\lambda - \exp(x/\bar\lambda)])$. The mean distance among points is calculated in two parts: $\langle y\rangle = \Psi(k)/c = (1/c)d[\ln\Gamma(k)]/dk$, that is digamma function [27], so that the mean distance among points is $\langle l^{(k)}_{\rho,d}\rangle = N_u^{-\beta}[1 + \beta\Psi(k)]$. The neighborhood order is an integer, which leads to a representation $\Psi(k) = -\gamma + \sum_{i=1}^{k-1} i^{-1}$, rewriting i^{-1} as $(k-i)^{-1}$, one has the mean distance on l:

$$\langle l^{(k)}_{\rho,d}\rangle = N_u^{-\beta}\left[1 + \beta\left(-\gamma + \sum_{i=1}^{k-1}\frac{1}{k-i}\right)\right], \tag{10}$$

where $\gamma = 0.57721\ldots$ is Euler's constant and for $k \gg 1$, $\Psi(k) \approx \ln(k)$ and $\langle l^{(k)}_{\rho,d}\rangle = N_u^{-\beta}[1 + \beta\ln(k)]$, this was firstly obtained by Cerf *et al.* conjecture expanding the term $\Gamma(k+\beta)/\Gamma(k)$ of Eq. 3. This term, on average, represent distance increment due to neighborhood order increase. Due to accurate approximation for ratio $\Gamma(z+x)/\Gamma(z)$, we demonstrate this conjecture using distance statistics. Furthermore, Eq. 9 allows us to calculate not only mean distance, but also variance and higher order moments.

The variance of Eq. 9 is $\sigma^2(y)_k = \Psi^{(1)}(k)$, where $\Psi^{(1)}(k)$ is trigamma function. One can argue that $\sigma^2(a+bx) = b^2\sigma^2(x)$, so that the standard deviation in l is

$$\sigma(l)_{\rho,d,k} = \frac{\beta N_u^{-\beta}}{\sqrt{k}}, \tag{11}$$

where we made use of the $\Psi^{(1)}(k) \approx 1/k$, for $k \gg 1$. In the l variable, the mean distance is only weakly affected by neighborhood order, and the variance vanishes rapidly. This occurs because, in high dimensionality, a small increase in the radius leads to a large increase in volume. The larger the radius is, the smaller is the increment to generate the same increase in volume. Therefore, the higher the neighborhood order, the smaller the radius increase is and the lower the standard deviation is around the mean value. The distance distribution in variable l is described as a delta sequence.

3.2 High neighborhood order

The second limiting case is the distance distribution for high neighborhood order. From Eq. 4, one sees that the standard deviation decays according to $k^{\beta-1/2}$, for $k \gg \beta$. According to the central limit theorem, the variance of the summation S, of N independent and identically distributed random variables, with finite variance, σ, is $\sigma_r = \sigma/\langle S\rangle$ and the skewness decreases as $1/\sqrt{N}$ [13]. For $k \gg 1$ and for arbitrary dimension, the relative standard deviation and skewness decrease as $1/\sqrt{k}$. This indicates that, besides recovering the symmetry of the pdf around its mean value, the neighborhood order increase makes Eq. 1 to converge to a Gaussian distribution. This behavior is obtained by numerical simulation and illustrated in the graphs of Figs. 1 and 2.

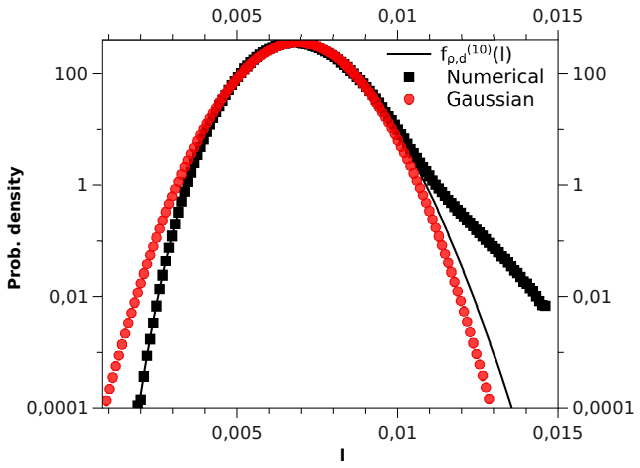

Fig. 2. Gaussian approximation for distance statistics for $k \gg 1$ and two dimensional media. Simulation parameters are $d = 2$, $\rho = 65365$ and $k = 10$. The weak adjustment on the tails is due the fact that the ends of the distribution converge more slowly than the peak.

The convergence to the Gaussian is due to the summation of volumes. The necessary volume around a point i to find k neighbors is on average kV_1, where V_1 is the volume needed to find the nearest neighbor, distance in this case is a random variable proportional to $(kV_1)^\beta$. Another way of understanding this convergence is considering the summation of the thicknesses of spherical shells comprising the same volume.

4 Applications

In this section, we discuss possible applications of our results in context of pseudo number generation and tests of departures from Poissonian hypothesis in spatial distributions. Table(1) enumerates various pdfs obtained by varying dimensionality of the medium and neighbor order in Eq. 1.

Due to the great amount of special cases of distance statistics, Table(1), one possible application is to use it as a very general pseudo random numbers generator. Although not efficient in terms of time consumption, it allows one to have a unified probability density functions arising from distance measurements in random media.

Table 1. Summary of probability distributions obtained from Eq. 1 for different dimensionalities and neighborhood orders. The symbol (-) means arbitrary value, ∞ is a high value and (*) means distribution in y variable given by Eq 7.

d	k	Distribution
1	1	Exponential
1	-	Gamma
1	∞	Gaussian
2	1	Rayleigh
2	-	Nakagami
3	-	Wilson-Hilferty
-	1	Weibull
-	-	Stacy
-	∞	Gaussian
∞	1	*Gumbel
∞	-	*Log-Gama
∞	∞	*Gaussian

Another possible application of Eq. 1 is to evaluate, whether given distances among points vary from Poissonian hypothesis. This evaluation was originally employed by Thompson [9], on the distance distribution among trees in a two dimensional environment. One way to assess deviations from this hypothesis is to perform a test of significance for the average distance to the kth nearest neighbor. The test uses limits of Eq. 1, when it is transformed into a χ^2 (chi square) distribution, Eq. 12. As a generalization of Thompson result, we propose the same test in an environment of arbitrary dimensionality. Rewriting Eq. 1 as a function of $x_n = 2\lambda$, it becomes:

$$f^{(k)}(x_n) = \frac{1}{2\Gamma(k)}\left(\frac{x_n}{2}\right)^{k-1}\exp(-x_n/2), \qquad (12)$$

that is χ^2 distribution, with $2k$ degrees of freedom. Once one knows the density of points on media, it is possible to apply the test and detect deviations in any neighborhood order, not only for the nearest one.

5 Conclusion

Using only Poisson process, we calculate the statistical distribution of distance for disordered media with arbitrary dimensionality. Our results have been validated by Monte Carlo simulations. Starting with Eq. 1 we calculate the limiting case of high dimensionality and high neighborhood order. Distance statistics on high dimensional case becomes a delta sequence around the mean distance, that was firstly conjectured by Cerf *et al.*. Distance statistics in high neighborhood order converges to a Gaussian distribution due to central limit theorem. The general pdf with $d < 3$ and arbitrary neighborhood order leads to special cases that retrieves well known pdfs such gamma, Weibull, etc. Distance statistics may detect departures from Poissoanian, as pointed by Thomson for $d = 2$, and generalized by Eq. 12, opening up new possibilities like three dimensional image analyzes of cells distribution, etc.

C.R.F.Granzotti acknowledges support from CAPES and FAP ESP(2010/00087-0). A.S.M. acknowledges support from CNP q(305738/2010-0 and 485155/2013-3) and CAPES. The authors would like to thank C.A.S Terariol, R. S. González and J. M. Berbert for useful discussions.

References

1. C. A. S. Terçariol, A. S. Martinez, Brazilian Journal of Physics **36**, 232 (2006)
2. P. J. Clark, F. C. Evans, Science **121**, 397 (1955)
3. P. J. Clark, Science **123**, 373 (1956)
4. M. F. Dacey, Geographical Analysis **1**, 385 (1969)
5. T. F. Cox, Biometrics **37**, 367 (1981)
6. C. A. S. Terçariol, F. d. M. Kiipper, A. S. Martinez, J. Phys. A: Math. Theor. **40**, 1981 (2007)
7. S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943)
8. P. J. Clark, F. C. Evans, Ecology **35**, 445 (1954)
9. H. R. Thompson, Ecology **37**, 391 (1956)
10. N. J. Cerf, J. Boutet de Monvel, O. Bohigas, O. C. Martin , A. G.Percus, J. Phys. I France **7**, 117 (1997)
11. A. Chakraborti, Int. J. Mod. Phys. C **12**, 857 (2001)
12. M. Mezard, G. Parisi, Europhysics Letters **49**, 2019 (1988)
13. J. Houdayer, J. Boutet de Monvel, O. Martin, Eur. Phys. J. B **6**, 383 (1998)
14. G. F. Lima, A. S. Martinez, O. Kinouchi, Phys. Rev. Lett. **87**, 010603 (2001)
15. C. A. S. Terçariol, R. S. González, A. S. Martinez, Phys. Rev. E **75**, 061117 (2007)
16. A. Tewari and A. Gokhale, Acta Materialia **54**, 1957 (2006)
17. T. Cover, P. Hart, IEEE Trans. Inf. Theory **13**, 21 (1967)
18. S. Singh, J. Haddon, M. Markou, Pattern Recognition **34**, 1601 (2001)
19. M. Haenggi, IEEE Trans. Inf. Theory **51**, 3584 (2005)
20. A. G. Percus, O. C. Martin, Adv. Appl. Math. **21**, 424 (1998)
21. P. Bhattacharyya, B. K. Chakrabarti, Eur. J. Phys. **29**, 639 (2008)
22. D. Evans, A. J. Jones, W. M. Schmidt, Proc. R. Soc. A **458**, 2839 (2002)
23. D. Evans, Proc. R. Soc. A **464**, 3175 (2008)
24. D. Moltchanov, Ad Hoc Networks **10**, 1146 (2012)
25. E. Stacy, Annals Mathematical Statistics **33**, 1187 (1962)
26. G.E. Crooks, arXiv:1005.3274 [math.ST], (2010)
27. M. Abramowitz, I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* (Dover books on mathematics, Dover, 1972)